
PrefLens: Position Artefacts and Convergent Validity in LLM Preference Elicitation ¹

Hafiza Hajrah Rehman
Saarland University

Ameesha
Saarland University

With
Apart Research

Abstract

Large language models can express apparently coherent preferences, but it is unclear whether different elicitation methods measure the same underlying construct. We introduce **PrefLens**, a framework for testing convergent validity and diagnosing measurement artefacts in LLM preference elicitation. Across three studies, we compared multiple elicitation procedures and examined sensitivity to option position, model scale, and prompt framing. Gemini showed strong cross-method convergence, while Qwen’s pairwise preference signal was fully explained by display position. For Llama, position adjustment increased self-report/pairwise correlation from $\rho = -0.007$ to $+0.524$, suggesting that order effects can also obscure agreement. In a provider-pinned comparison, GPT-OSS 120B was more position-susceptible than 20B, and this persisted after removing an explicit indifference cue. These findings show that apparent LLM preferences can depend strongly on measurement design. We recommend exact order counterbalancing and reporting position-conditioned choice rates alongside aggregate preference estimates.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

1. Introduction

Large language models often produce stable-looking choices, motivating attempts to measure their apparent preferences and values. However, there is **no independent ground truth for what an LLM “prefers.”** A measured preference may instead reflect prompt wording, response heuristics, or option order. This makes preference elicitation fundamentally a measurement-validity problem. We study this through convergent validity: if different elicitation methods recover similar preference rankings, the measurement is more robust; disagreement may instead expose method-specific artefacts. We focus especially on display position, which prior work shows can strongly affect LLM choices.

We introduce PrefLens, a framework for comparing preference-elicitation methods while diagnosing positional effects across three studies.

Our main contributions are:

1. A multi-method convergence test using self-report, pairwise choice, cost trade-offs, and sequential task selection.
2. A position/content decomposition showing that option order can create apparent preference signal and can also obscure cross-method agreement.
3. Controlled follow-ups showing that GPT-OSS 120B was more position-sensitive than 20B, and remained strongly position-dominated after removing an explicit indifference cue.

We interpret these findings as properties of elicited choice behaviour, not evidence of genuine internal preferences or consciousness.

2. Related Work

The closest work on LLM preference measurement is Utility Engineering (Mazeika et al., 2025), which repeatedly samples pairwise choices and fits utility models to test whether LLM preferences exhibit coherent structure. It reports increasing utility coherence with model scale. PrefLens builds on the general idea of repeated behavioural preference elicitation but asks a different question: **do different operationalisations of preference recover the same structure, and what nuisance variables explain disagreement when they do not?**

Separately, position and selection biases are well established in LLM evaluation. Pezeshkpour and Hruschka show large performance changes when multiple-choice options are reordered, while subsequent work demonstrates position bias in LLM-as-a-judge and other ranking settings. These studies establish that order can influence LLM decisions. Our work connects this failure mode specifically to preference measurement: position bias can constitute the measured preference signal itself, and removing realised order imbalance can change conclusions about cross-method agreement. PrefLens therefore complements prior utility and position-bias work by treating preference elicitation explicitly as a measurement-validity problem rather than relying on any single elicitation procedure.

3. Methods

3.1 Study 1: Multi-method preference elicitation

We tested 12 welfare-neutral task pairs plus two sanity controls across Llama-3.1-8B-Instruct, Qwen-2.5-7B-Instruct, and Gemini-3.1-Flash-Lite. Sampling used temperature 1.0 and top-p 1.0. All prompts, responses, choices, display orders, failures, and metadata were logged, yielding 4,013 trial records. Because Gemini did not complete Method D, cross-model comparisons use the common complete subset of three methods $\times$ ten items.

We measured apparent preference using four procedures, each mapped to $[-1, +1]$:

Table 1. Preference-elicitation procedures used in Study 1.

Method	Procedure	Measure
Self-report	Choice + preference strength	mean signed strength
Pairwise	Repeated A/B choices	$(2P(A)-1)$
Cost trade-off	Choice under symmetric added cost	mean across cost levels
Sequential	Repeated task selection	signed occupancy

Method C was originally designed around an indifference point, but a manipulation check showed non-monotonic cost responses, so we used mean choice rate across cost levels instead. Display order was randomised and converted back to semantic A/B coordinates before scoring. Convergence was evaluated using direction agreement and Spearman correlation, with a 10,000-permutation matched null that preserved each method’s score distribution while breaking item alignment. We also separated position from order-invariant content:

$$position = P(X|X \text{ first}) - P(X|X \text{ second}), \quad content = P(X|X \text{ first}) + P(X|X \text{ second}) - 1$$

The content term represents an order-invariant choice signal, not a genuine internal preference.

3.2 Studies 2b and 3: Controlled position tests

To isolate position effects within one model family, we compared **GPT-OSS 20B and 120B** using the same Groq provider. Each semantic option appeared first and second exactly ten times per item, producing **480 preference and 80 control trials**. Before data collection, we predicted that 120B would show **less position sensitivity (H1)** and **more order-invariant content signal (H2)**. Confidence intervals used 10,000 bootstrap resamples over items.

Study 3 repeated the same design but removed one sentence stating that neither choice was more correct or helpful, testing whether this explicit indifference cue contributed to positional responding.

4. Results

4.1 Cross-method convergence was strongly model-dependent

On the matched three-method, ten-item comparison, Gemini showed high agreement across elicitation procedures, whereas Llama was intermediate and Qwen showed little cross-method alignment.

Table 2. Study-1 convergence on the matched three-method $\times$ ten-item subset.

Model	Direction agreement	Mean Spearman ρ	Permutation p , agreement	Permutation p , ρ
Gemini-3.1-Flash-Lite	0.911	+0.868	0.0019	0.0001
Llama-3.1-8B	0.690	+0.308	0.123	0.069
Qwen-2.5-7B	0.565	+0.021	0.274	0.433

Thus, only Gemini clearly exceeded its matched permutation null on the common comparison basis. Llama is more ambiguous: on its fuller four-method dataset its convergence is nominally significant, but this does not survive correction for multiple comparisons. We therefore treat Llama as suggestive rather than established.

4.2 Position could manufacture (and obscure) apparent preference structure

For Qwen, pairwise choices were entirely explained by display position: across all 12 items, semantic-A choice frequency matched how often A appeared first, and the order-adjusted content signal was 0.000 for every item. Sanity controls were still answered correctly, ruling out a global task-comprehension failure. Position could also obscure agreement. In an exploratory self-report/pairwise reanalysis, equal-weighting display orders increased Llama’s correlation from $\rho = -0.007$ to $+0.524$ and direction agreement from 0.571 to 0.833, consistent with order imbalance contributing to apparent disagreement. For Qwen, adjustment reduced the pairwise vector to a constant zero, making rank correlation undefined.

4.3 Larger scale did not eliminate positional susceptibility

The controlled GPT-OSS experiment rejected both pre-specified hypotheses. In the provider-pinned replication, mean absolute position effect was 0.608 for 20B and 0.942 for 120B. The paired difference,

$$\Delta|position|_{20B-120B} = -0.333 \quad [95\% CI: -0.492, -0.192],$$

was opposite to H1. Likewise, mean absolute content signal was 0.208 for 20B and 0.058 for 120B, giving

$$\Delta|content|_{120B-20B} = -0.150 \quad [95\% CI: -0.217, -0.092]$$

opposite to H2. Thus, within this family and serving provider, the larger model was **more position-susceptible**, not less. This does not establish that scale causes position bias. Model sizes can still differ in training data, compute and post-training but it rejects the simple account that positional responding is merely a small-model incapacity.

The sanity controls produced a particularly useful dissociation: both provider-pinned GPT-OSS models scored 100% control accuracy, while 120B remained almost completely position-driven on the balanced items. The model therefore could override position when the semantic distinction was decisive, ruling out a global prompt-reading or option-parsing failure.

4.4 Removing the explicit indifference cue did not remove the 120B effect

Study 3 removed only the explicit statement that neither choice was better. GPT-OSS 120B remained extremely position-dominated: mean $|\text{position}|$ changed only from 0.942 to 0.925 ($\Delta=+0.017$, 95% CI $[-0.025, +0.067]$). We did not pre-specify an equivalence margin, so this is not evidence that the two framings are identical; the narrower conclusion is that extreme positional responding persisted without the explicit indifference cue.

For 20B, removing the cue reduced $|\text{position}|$ from 0.608 to 0.483 ($\Delta=+0.125$, 95% CI $[+0.025, +0.225]$). The model-by-framing interaction was -0.108 $[-0.208, -0.008]$. However, these effects do not survive a strict correction across the five reported framing tests, so we treat them as suggestive evidence that elicitation wording may modulate positional susceptibility differently across models, rather than as established effects.

5. Discussion and Limitations

Our results show that LLM preference measurement is highly sensitive to the elicitation procedure. Position could create an apparent preference signal in Qwen, while removing position increased Llama’s self-report/pairwise agreement. In GPT-OSS, the larger 120B model was more position-sensitive than 20B, and this remained true after removing the explicit indifference cue.

For AI safety, this means apparent model preferences should not be interpreted without checking for nuisance effects such as option order. Cross-method agreement is useful, but only when the measurements themselves are robust.

Limitations

We have no ground truth for genuine model preferences. The study uses only 12 items and a small set of models, Gemini lacks one method, and the position-adjusted analysis covers only self-report and pairwise choice. Studies 2b and 3 test only pairwise choice, and model scale remains confounded with training differences.

Future Work

Future work should test more models and preference domains, use exact counterbalancing across all elicitation methods, and study interactions between position, framing, and sampling settings.

6. Conclusion

PrefLens shows that position can create apparent preference signal, obscure agreement between methods, and affect models differently. We therefore recommend exactly counterbalancing option order and reporting position-conditioned choice rates alongside aggregate preference estimates.

These findings do not identify true model preferences; they show why stronger measurement controls are needed before making such interpretations.

Code and Data

- **Code repository:** github.com/hajraRehman/PrefLens
- **Data and results:** Included in the repository under data/ and results/.
- **Reproducibility:** Configurations, raw outputs, analysis code, tests, and decision logs are included.

References

1. Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Zhang, O., & Hendrycks, D. (2025). **Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs.** *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. [Paper](#)
2. Pezeshkpour, P., & Hruschka, E. (2024). **Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions.** *Findings of the Association for Computational Linguistics: NAACL 2024*, 2006–2017. DOI: 10.18653/v1/2024.findings-naacl.130. [ACL Anthology](#)
3. Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). **Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.** *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*. DOI: 10.52202/075280-2020. [NeurIPS paper](#)

Appendix

Appendix A. Experimental Details

Study 1 used 12 welfare-neutral preference pairs and two sanity controls. All methods mapped responses into semantic A/B coordinates so that display position could be analysed separately from option identity. Raw prompts, model responses, parsing status, sampling parameters, display orders, and experiment configurations are included in the public repository.

Study 2b and Study 3 used exact counterbalancing: for every model–item pair, semantic option X appeared first ten times and second ten times. GPT-OSS 20B and 120B were served through the same Groq upstream provider, which was recorded and verified for every response. Study 3 was identical to Study 2b except for removal of the sentence stating that neither option was more correct or helpful.

All uncertainty intervals reported for the GPT-OSS comparisons were obtained by bootstrapping preference **items**, rather than treating individual API responses as independent observations.

Appendix B. Additional Analyses and Robustness Checks

The matched Study-1 permutation null preserved each method's observed score distribution while independently permuting item identities across method columns. This retains ties, saturation, and method-specific response distributions while removing cross-method item alignment.

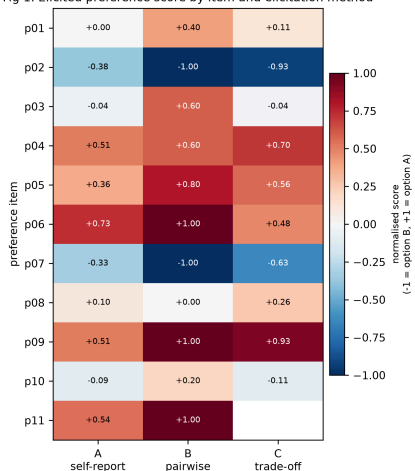
Direction-agreement conclusions were also evaluated across multiple dead-zone thresholds. Gemini remained the strongest-converging model across the tested thresholds, although the relative ordering of Llama and Qwen was not invariant.

The exploratory position-adjusted analysis was restricted to self-report and pairwise choice. Exact adjustment could not be reconstructed for the trade-off procedure because 22.7% of item $\times$ cost-level cells lacked one display order, while sequential-task scores aggregate stages with independently randomised orders.

Appendix C. Additional Figures

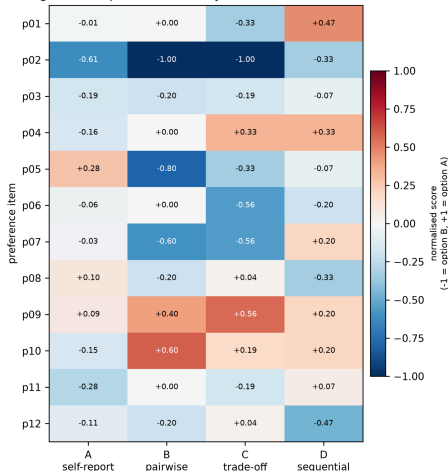
Figure C1: Study 1 preference heatmaps

Fig 1. Elicited preference score by item and elicitation method

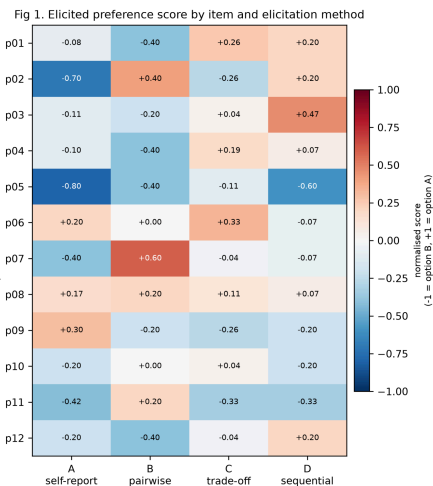


Gemini

Fig 1. Elicited preference score by item and elicitation method



Llama

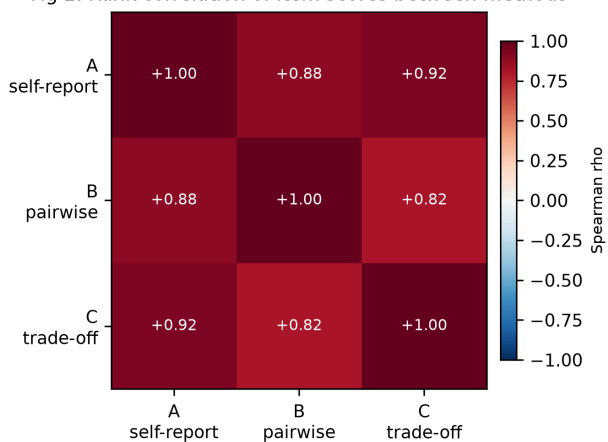


Qwen

Figure C1. Preference scores across elicitation methods for Gemini, Llama, and Qwen. Rows correspond to preference items and columns to elicitation procedures. The figure illustrates strong cross-method structure for Gemini, weaker alignment for Llama, and highly inconsistent measurements for Qwen.

Figure C2: Method correlation matrices

Fig 2. Rank correlation of item scores between methods



model: gemini-31-flash-lite | phase: main | 11 preference items | scores in [-1,+1], + favours semantic option A

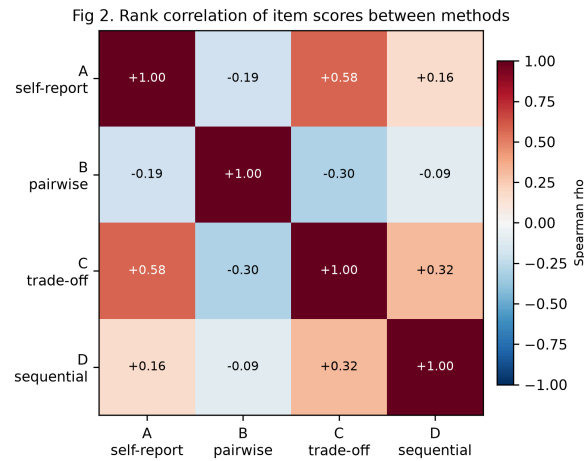
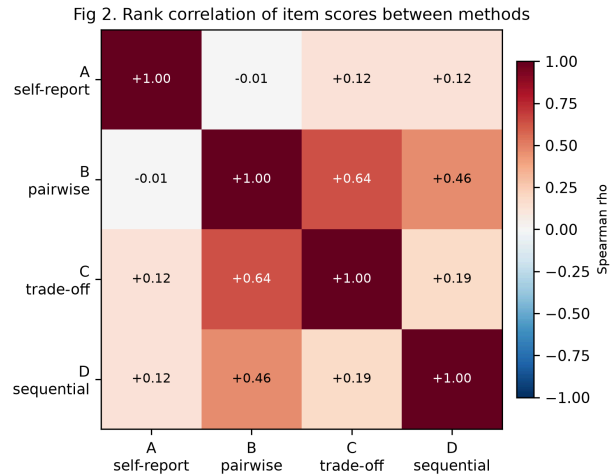


Figure C2. Pairwise correlations between preference-elicitation procedures in Study 1. The matrices show substantial variation in cross-method agreement across models, motivating the matched permutation analysis and subsequent position diagnostics.

Appendix D. Limitations and Dual-Use / Ethical Considerations

Limitations

PrefLens measures elicited choice behaviour rather than genuine internal preferences, for which no independent ground truth is available. Results are based on a limited set of models and preference items, and the GPT-OSS scale comparison does not isolate model size causally.

Dual-use and ethical considerations

More reliable preference elicitation may improve research on model values and welfare-relevant behaviour, but could also facilitate behavioural steering or manipulation. Cross-method consistency should not be interpreted as evidence of consciousness, subjective experience, or moral status.

LLM Usage Statement

We used LLM assistance for brainstorming, implementation support, and language editing. All experimental results, statistical analyses, and claims in the final report were checked against the committed data and analysis outputs by the author.